UniProt/PIR Use Cases

Baris E. Suzek
Hongzhan Huang

Use Case 1: Find protein information

CHARACTERISTIC INFORMATION

Goal in Context: To find protein information in UniProt/PIR data source

Preconditions:

1. UniProt/PIR web service is advertised as a protein data source in caGRID system (Use case 3.4 from caGRID white paper).
2. UniProt/PIR data source and UniProt/PIR object mappings are configured at user site (Use case 2.2 and 2.3 from caGRID white paper).
3. A caGRID query client is available.

Success End Condition: The user finds information on proteins of interest.

Failed End Condition:

1. The user is unable to find any information on proteins of interest.
2. Loss of network connection before retrieval of the results is completed.

Primary Actor: Researcher, Scientist

MAIN SUCCESS SCENARIO

1. The user sets search criteria for the proteins of interest (Use case 2).
2. The user sets response criteria (Use case 3).
3. The user invokes the query using caGRID client.
4. UniProt/PIR web service returns the results to caGRID client in UniProt XML format.

Use Case 2: Setting search criteria

CHARACTERISTIC INFORMATION

Goal in Context: To set search criteria to retrieve proteins of interest

Preconditions:
1. caGRID query client is ready to submit a query to UniProt/PIR web service
2. caGRID query client provides an interface to formulate a query

Success End Condition: The user can set search criteria using the fields provided by UniProt/PIR web service

Failed End Condition: The user is unable to set search criteria for the proteins of interest.

Primary Actor: Researcher, Scientist

MAIN SUCCESS SCENARIO

1. The user formulates a query by using
- an individual field (simple search criteria) or
- fields combined with Boolean operators such as "AND" , "OR" and "NOT" (advanced search criteria) or
- an "all fields" wildcard to invoke a google-like search for all the available fields (text search criteria)

The fields available in UniProt/PIR data source are:
- UniProt ID or accession number
- Protein name
- Gene name or symbol
- Keywords
- Scientific or common organism name
- NCBI Taxonomy ID
- PIR ID or accession number
- NCBI GI ID
- GenPept accession number
- Refseq accession number
- Locus ID
- PDB ID with/without chain ID
- OMIM ID
- TIGR ID
- EMBL ID
- UniRef100/90/50 ID
- UniParc ID
- PubMed ID
- PIRSF ID
- PFAM ID
- EC number

- PROSITE ID
- PRINTS ID
- GO ID
- InterPro ID
- TIGRFAMS ID
- Sequence length (with range operators such as “<”, “>”,”>=” etc.)
- Molecular weight (with range operators such as “<”, “>”,”>=” etc.)

Use Case 3: Setting Response Criteria

CHARACTERISTIC INFORMATION

Goal in Context: To set response criteria for the results returned by UniProt/PIR web service

Preconditions:

1. caGRID query client is ready to submit a query to UniProt/PIR web service.
2. caGRID query client provides an interface to set the criteria on the results returned by UniProt/PIR web service.

Success End Condition: The user sets response criteria

Failed End Condition: The user is unable to find the response options that cover the information needed.

Primary Actor: Researcher, Scientist

MAIN SUCCESS SCENARIO

1. The user sets the information content of the result returned by UniProt/PIR web service. The user may choose to use the information provided by default response(in UniProt XML format) , in which each protein record contains:
    - UniProt ID and accession number(s)
    - Protein name(s) and components
    - Gene name(s) and symbol(s)
    - Keywords
    - Scientific and common organism name
    - NCBI Taxonomy ID
    - Primary reference(literature citation)
    - PIR ID and accession number
    - NCBI GI ID
    - Locus ID
    - PDB ID with/without chain ID
    - PIR SuperFamily ID
    - Sequence length
    - Sequence checksum (CRC64)
    - Molecular weight
    - Sequence

    The user can also add one or more of the following to the default response:
    - Gene location (e.g. plasmid)
    - Comments (free text comments on the entry such as description of an enzyme regulatory mechanism or tissue specificity)
    - Features (annotation of the sequence data. e.g. splice variants, active sites)
    - All the references(literature citations)
    - The database IDs or accessions not included in the default set

2. The user sets the maximum number of protein records to returned by the UniProt/PIR web service. The user may choose to use the default value or override it.